BINER
Jurnal Ilmiah Informatika dan Komputer
Vol. 4 | No. 2 | Juli 2025

p-ISSN : 2828-0210 | e-ISSN : 2828-0229
https://ojs.unsiq.ac.id/index.php/biner

APLIKASI PEMBELAJARAN TAJWID BERBASIS SPEECH RECOGNITION MENGGUNAKAN METODE CNN-LSTM

Elfan Rizqi Awaludin [1)], Hidayatus Sibyan [2)], Nur Hasanah [3)]
[1,2,3)] Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer
Email: elfanriskiawaludin@gmail.com [1)]

Diterima : 10 Juli 2025 ; Disetujui : 30 Juli 2025 ; Dipublikasikan : 31 Juli 2025

ABSTRAK

Penelitian ini bertujuan untuk membangun sistem deteksi bacaan tajwid dengan menggunakan pendekatan deep learning berbasis speech recognition dengan arsitektur Convolutional Neural Network dan Long Short-Term Memory (CNN-LSTM). Sistem ini dirancang untuk mendeteksi apakah bacaan yang dilakukan pengguna termasuk "benar" atau "salah" berdasarkan karakteristik fitur suara yang diperoleh melalui ekstraksi Mel-Frequency Cepstral Coefficients (MFCC). Dataset yang digunakan terdiri dari lebih dari 350 file audio bacaan ayat pendek, kemudian diperluas menjadi 700 dataset melalui teknik augmentasi dan denoise untuk meningkatkan variasi dan kualitas data. Model CNN-LSTM dilatih menggunakan data MFCC berdimensi tetap dan dievaluasi berdasarkan tingkat akurasi serta confusion matrix pada data validasi. Hasil penelitian menunjukkan bahwa model memiliki performa yang baik dalam mengklasifikasikan bacaan dengan tingkat akurasi yang cukup baik. Evaluasi menggunakan uji model memberikan hasil akurasi pelatihan sebesar 90,83% dan akurasi validasi sebesar 77,78% yang bisa dibilang cukup akurat. Sistem ini diintegrasikan ke dalam aplikasi berbasis Flask dan Flutter, sehingga pengguna dapat mendapatkan hasil klasifikasi secara real-time. Diharapkan penelitian ini dapat menjadi kontribusi dalam pengembangan media pembelajaran Al-Qur'an yang adaptif dan berbasis teknologi suara.

Kata Kunci : Deteksi bacaan, tajwid, deep learning, CNN-LSTM, pengenalan suara, MFCC.

ABSTRACT

This study aims to build a tajweed recitation detection system using a deep learning approach based on speech recognition with a Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) architecture. This system is designed to detect whether the user's recitation is "correct" or "incorrect" based on the characteristics of the sound features obtained through the extraction of Mel-Frequency Cepstral Coefficients (MFCC). The dataset used consists of more than 350 audio files of short verse recitations, then expanded to 700 datasets through augmentation and denoising techniques to improve data variety and quality. The CNN-LSTM model was trained using fixed-dimensional MFCC data and evaluated based on the accuracy level and confusion matrix on the validation data. The results showed that the model has good performance in classifying recitations with a fairly good level of accuracy. Evaluation using a model test resulted in a training accuracy of 90.83% and a validation accuracy of 77.78%, which can be considered quite accurate. This system is integrated into Flask and Flutter-based applications, so users can get classification results in real-time. It is hoped that this research can contribute to the development of adaptive Al-Qur'an learning media based on sound technology.

Keywords: *Reading detection, tajweed, deep learning, CNN-LSTM, speech recognition, MFCC.*

1. PENDAHULUAN

Ilmu tajwid merupakan ilmu paling utama yang wajib diketahui oleh setiap muslim. Ilmu tajwid merupakan ilmu yang mempelajari cara membaca Al-Qur'an yang benar dan tepat. Namun masih banyak sekali masyarakat yang kesulitan dalam membaca Al-Qur'an secara baik dan benar. Bahkan masih ada yang masih buta huruf Al-Qur'an. Berdasarkan riset PTIQ Jakarta, umat Islam Indonesia yang tidak bisa membaca Al-Qur'an ada sekitar 60- 70%. Kalau dibuat ringkasan dari temuan-temuan itu, kurang lebih ada 50-60% umat Islam belum bisa membaca Al-Qur'an [1].

Hal ini dikarenakan belum menemukan metode yang cepat dan mudah untuk belajar membaca Al-Qur'an sehingga mereka malas untuk belajar. Dalam membantu pembaca untuk bisa membaca Al-Qur'an secara baik dan benar dari dasar, tentunya harus ada pengembangan pembelajaran dasar dalam ilmu tajwid itu sendiri yang lebih interaktif. Beberapa kendala yang muncul dalam sistem pembelajaran tajwid saat ini antara lain: (1) Kurangnya media bantu interaktif berbasis suara dan media digital untuk mengenali kesalahan pelafalan bacaan. (2) Tidak adanya umpan balik langsung terhadap bacaan, (3) Minimnya penggunaan teknologi digital untuk menunjang efektivitas pembelajaran tajwid secara mandiri, (4) Rasa malas seseorang dalam mempelajari ilmu tajwid dalam media cetak. kurangnya Metode Pembelajaran yang *inovatif* dan interaktif, metode pembelajaran konvensional sering kali dianggap kurang menarik oleh generasi muda yang lebih terbiasa dengan teknologi digital. Minimnya visualisasi dan pengalaman langsung dalam memahami ilmu tajwid membuat pembelajaran menjadi kurang efektif.

Dalam mempelajari ilmu tajwid, tentunya tidak sedikit waktu untuk bisa mempelajarinya secara teliti dan benar. Untuk itu harus ada inovasi terbaru dalam mengikuti berkembangnya zaman agar ilmu agama tidak lenyap seketika. Salah satu perkembangan aplikasi yang memberikan solusi untuk penggunanya dalam belajar yaitu aplikasi interaktif dan edukatif dalam ilmu tajwid. Untuk mendukung pengembangan aplikasi tersebut, peneliti menggabungkan fitur mendeteksi bacaan yaitu dengan pengenalan suara atau *speech recognition*. *Speech recognition* telah banyak diaplikasikan ke dalam sistem seperti pada penelitian [2] yang menggunakan *speech recognition* sebagai alat bantu komunikasi untuk tunawicara.[3].

Untuk mendorong terciptanya solusi dari permasalahan tersebut, peneliti memakai beberapa algoritma yang dapat digunakan dalam perkembangan *speech Recognition* dalam pembuatan aplikasi Ilmu Tajwid seperti CNN dan LSTM. Salah satu algoritma yang dirancang untuk menangani data berurutan seperti suara adalah *long short-term memory* (LSTM) [4]. LSTM adalah jenis khusus dari arsitektur jaringan saraf tiruan berulang *(recurrent neural network)* yang dirancang untuk mengatasi masalah ketergantungan jangka panjang dan jangka pendek dalam urutan data [5]. LSTM memiliki konsep gate yang mengatur aliran informasi yang masuk dan keluar dari setiap neuron dalam layer LSTM.

Tujuan penelitian ini adalah (1) Mengembangkan sistem dan menerapkan model CNN-LSTM pembelajaran berupa Speech Recognition yang dapat mendeteksi dan mengevaluasi bacaan meningkatkan akurasi deteksi kesalahan dalam bacaan tajwid dan huruf hijaiyah dengan benar dalam kaidah Tajwid. (2) Mengukur tingkat akurasi dan evektivitas sistem pembelajaran Tajwid berbasis Speech Recognition dengan metode CNN-LSTM. Penulis membuat dan mengambangkan Aplikasi pembelajaran Tajwid dengan teknologi *Speech Recognition* menggunakan metode CNN-LSTM pada platform android, dengan harapan agar dapat mengevaluasi pengalaman pengguna terhadap penggunaan Aplikasi sebagai penggabungan teknologi modern dengan praktik keagamaan.

2. METODE

Penelitian ini menerapkan metode kuantitatif eksperimen dengan rekayasa sistem yang berlandaskan deep learning dan dalam bidang pengenalan suara (speech recognition), khususnya pada cabang klasifikasi langsung terhadap sinyal audio (direct audio classification). Sistem yang dikembangkan bertujuan untuk mengidentifikasi bacaan tajwid dari pengguna sebagai "benar" atau "salah" berdasarkan analisis terhadap pola akustik suara, tanpa melalui proses konversi ke dalam bentuk teks (tanpa automatic speech-to-text).

Desain penelitian ini bertujuan untuk membangun model deteksi suara bacaan tajwid yang dapat menentukan kualitas bacaan sebagai "benar" atau "salah" berdasarkan sifat fonetik yang ada. Model ini dikembangkan menggunakan arsitektur CNN-LSTM (Convolutional Neural Network – Long Short-Term Memory), yang dipilih karena kemampuannya dalam menggabungkan ekstraksi fitur spasial dari CNN dan analisis urutan temporal dari LSTM.

Fokus penelitian ini adalah pada data suara bacaan tajwid dari ayat-ayat pendek Al-Qur'an, yang diucapkan pengguna dengan beragam latar belakang serta pelafalan. Dataset yang digunakan terdiri dari lebih dari 350 rekaman suara, yang kemudian diduplikasi dengan teknik augmentasi suara dan pengurangan noise untuk meningkatkan variasi data serta memperluas kemampuan generalisasi model. Proses pengumpulan data dilakukan melalui rekaman langsung dan pengumpulan file audio dalam format .wav, dengan kriteria standar: mono channel, sample rate 16.000 Hz, dan durasi 1-3 detik. Semua data dilabeli secara manual ke dalam dua kategori, yaitu "benar" dan "salah", berdasarkn aturan ilmu tajwid.

Arsitektur utama dalam penelitian ini adalah model CNN-LSTM yang dikembangkan dengan menggunakan pustaka TensorFlow dan Keras. Proses ekstraksi fitur suara dilakukan menggunakan metode MFCC (Mel-Frequency Cepstral Coefficients), yang terbukti efektif dalam menggambarkan ciri khas suara manusia. Pelatihan dan pengujian model dilakukan menggunakan Google Colaboratory dengan dukungan GPU yang tersedia secara gratis, sedangkan pengolahan data dilakukan dengan bahasa Python menggunakan pustaka pendukung seperti Librosa, Pandas, dan NumPy. Untuk kebutuhan integrasi sistem, hasil model dikemas dalam format .h5 dan diimplementasikan dalam sistem backend menggunakan framework Flask. Aplikasi untuk pengguna dibangun menggunakan Flutter, yang menghubungkan hasil rekaman suara ke server Flask dan menampilkan hasil klasifikasi secara langsung.

Analisis data dilakukan dengan mengevaluasi performa model menggunakan metrik akurasi, confusion matrix, serta visualisasi kurva akurasi dan loss selama pelatihan. Selain itu, untuk menguji sistem dalam pengenalan suara, juga menggunakan BlackBox. Dengan pendekatan ini, diharapkan sistem yang dikembangkan tidak hanya tepat dalam mengenali hukum tajwid, tapi juga memberikan umpan balik edukatif yang bermanfaat bagi pengguna. Validitas sistem dipastikan melalui uji coba berulang pada data yang belum dilatih, serta verifikasi manual terhadap hasil deteksi.

3. HASIL DAN PEMBAHASAN

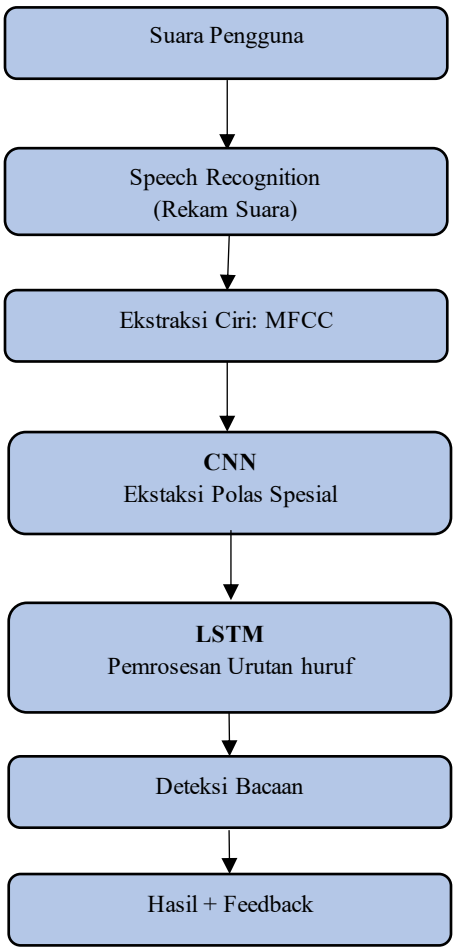
Penelitian ini menghasilkan sebuah sistem atau aplikasi pembelajaran tajwid yang dilengkapi dengan fitur input suara dari pengguna dan dilengkapi feedback bacaan secara realtime. Sistem ini mengimplementasikan pendekatan *direct audio classification* dalam ranah *speech recognition*, di mana sinyal suara yang masuk tidak dikonversi menjadi teks, melainkan langsung dianalisis sebagai bentuk audio mentah menggunakan fitur MFCC (Mel-Frequency Cepstral Coefficients). Proses pelatihan dilakukan pada data berdimensi tetap sebesar 160 frame waktu dan 40 koefisien MFCC per frame, menghasilkan input berbentuk matriks (160, 40) untuk setiap sampel suara.

Model CNN-LSTM yang digunakan terbukti mampu mengenali pola spasial dan temporal dari fitur suara. CNN berperan dalam mengekstrak pola lokal dari spektrum suara, seperti ciri pelafalan dengung, qalqalah, atau tekanan bacaan, sedangkan LSTM menangkap urutan bunyi dan dinamika waktu dalam bacaan yang dapat membedakan antara pelafalan tajwid yang tepat dan yang salah.

3.1. Desain Sistem Pembelajaran Tajwid

Sistem ini terdiri dari dua bagian utama: model klasifikasi berbasis deep learning dan sistem antarmuka pengguna. Model dibangun dengan menggunakan arsitektur CNN-LSTM, di mana CNN bertugas mengekstrak pola spasial dari fitur MFCC, dan LSTM mempelajari urutan waktu dari pelafalan pengguna. Dataset audio yang digunakan berasal dari 300+ file suara bacaan ayat pendek, kemudian diproses menjadi bentuk MFCC dengan parameter n_mfcc = 40, sr = 16000, dan panjang 1–5 detik. Untuk standarisasi dimensi input model, digunakan padding/trimming agar semua input menjadi (160, 40). Model diimplementasikan dalam

TensorFlow dan dikemas dalam format .h5 untuk kemudian diintegrasikan dengan server Flask dan aplikasi Flutter.



Gambar 1 Analisis Data

3.2. Proses Algoritma CNN-LSTM

Proses dimulai dari pengumpulan dataset audio yang direkam secara manual dari beberapa penutur dengan variasi pelafalan. Dataset disusun dalam format .wav mono 16-bit dengan sampling rate 16 kHz dan dikategorikan secara seimbang dalam dua kelas (benar/salah),

Tabel 1 Dataset .csv

Filename	Label	Teks
samiiun1.wav	benar	سَمِيعٌ بَصِيرٌ
minhasad1.wav	benar	مِنْ حَسَدٍ
minhasad1_denoise.wav	benar	مِنْ حَسَدٍ
minhasad2.wav	benar	مِنْ حَسَدٍ
samiiun_salah1.wav	salah	سَمِيعٌ بَصِيرٌ
samiiun_salah2.wav	salah	سَمِيعٌ بَصِيرٌ

kemudian dicatat dalam file dataset.csv Setiap file audio melalui tahap preprocessing, termasuk padding/potong durasi agar konsisten (±4 detik), lalu diekstraksi menggunakan fitur **MFCC** sebanyak 40 koefisien per frame dengan panjang 1600 frame waktu. Hasil ekstraksi ini menjadi input untuk model CNN-LSTM berdimensi (160, 40).

Model CNN-LSTM ini dirancang untuk mengklasifikasikan bacaan tajwid menjadi dua kategori, yaitu "benar" dan "salah", berdasar pada fitur audio MFCC yang memiliki durasi antara 1 hingga 4 detik. Arsitektur model ini dimulai dengan dua lapisan konvolusi 1D yang masing-masing diikuti oleh max pooling, normalisasi batch, serta dropout, bertujuan untuk mengekstrak dan menstabilkan fitur lokal dari sinyal audio.Setelah fitur spasial berhasil diperoleh, layer LSTM dengan 128 unit digunakan untuk menangkap pola temporal atau urutan waktu dari fitur suara, yang sangat penting dalam memahami urutan pengucapan huruf.Output dari LSTM kemudian dikendalikan melalui dua dense layer untuk memproduksi prediksi akhir mengenai dua kategori dengan menggunakan fungsi aktivasi softmax.Dengan total parameter sekitar 178 ribu, model ini cukup ringan dan efisien untuk diterapkan dalam aplikasi mobile, serta memiliki mekanisme perlindungan terhadap overfitting melalui penerapan dropout dan normalisasi batch.

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv1d_2 (Conv1D)	(None, 156, 64)	12,864
max_pooling1d_2 (MaxPooling1D)	(None, 78, 64)	0
batch_normalization_2 (BatchNormalization)	(None, 78, 64)	256
dropout_3 (Dropout)	(None, 78, 64)	0
conv1d_3 (Conv1D)	(None, 76, 128)	24,704
max_pooling1d_3 (MaxPooling1D)	(None, 38, 128)	0
batch_normalization_3 (BatchNormalization)	(None, 38, 128)	512
dropout_4 (Dropout)	(None, 38, 128)	0
lstm_1 (LSTM)	(None, 128)	131,584
dropout_5 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8,256
dense_3 (Dense)	(None, 2)	130

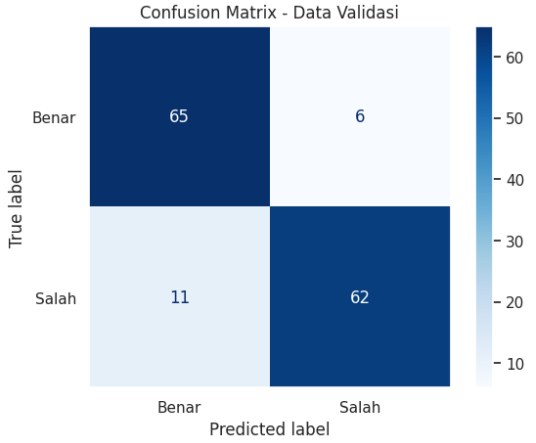
Total params: 178,306 (696.51 KB)
Trainable params: 177,922 (695.01 KB)
Non-trainable params: 384 (1.50 KB)

Gambar 2 Arsitektur Model *Cnn-Lstm*.

3.3. Hasil Pelatihan

Model dilatih menggunakan data latih dan validasi dengan rasio 80:20. Proses pelatihan dilakukan selama 30 epoch dengan fungsi aktivasi softmax, optimizer Adam, dan loss function categorical_crossentropy. Grafik pelatihan disajikan pada Gambar 3.1 yang

menunjukkan peningkatan akurasi training hingga 90,83%, dan akurasi validasi mencapai 77,78%. Meskipun terdapat sedikit gap akurasi, grafik menunjukkan proses pembelajaran berlangsung stabil tanpa fluktuasi besar, menandakan model tidak mengalami overfitting berlebihan.



Gambar 3 Confusion matrix

Confusion matrix disajikan pada Gambar 3.3, memperlihatkan bahwa sebagian besar prediksi "benar" dan "salah" diklasifikasikan dengan tepat.

3.5. Intregrasi Sistem ke Flask

Model yang telah dilatih disimpan dalam server Flask dan dihubungkan ke antarmuka Flutter. Pengguna dapat merekam suara melalui aplikasi Flutter, yang kemudian dikirim ke backend Flask dalam format .wav mono 16 kHz. Server akan mengekstrak fitur MFCC dan menjalankan prediksi menggunakan model .h5, lalu mengirimkan kembali label klasifikasi dan nilai confidence ke aplikasi. Umpan balik ditampilkan dalam bentuk teks dan animasi Unity yang sesuai dengan hasil klasifikasi.

3.6 Respon Sistem dan Akurasi Waktu Nyata

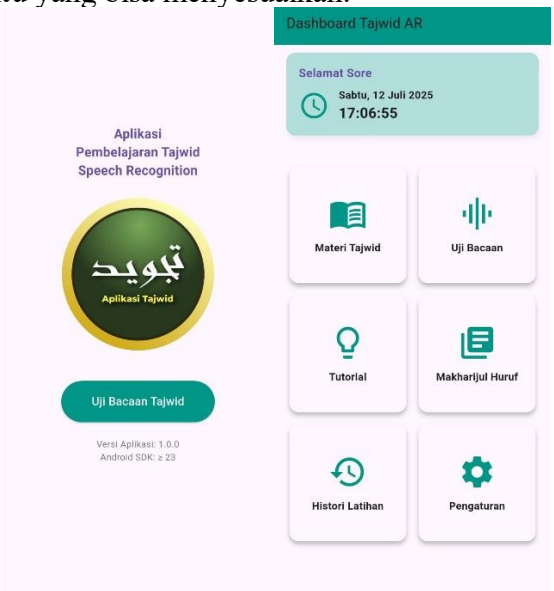
Sistem diuji dalam skenario pengguna asli, dan mampu memberikan hasil klasifikasi dalam waktu kurang dari 2 detik sejak audio dikirim. Hasil klasifikasi ditampilkan dalam UI secara real-time. Tabel 2 menyajikan beberapa contoh hasil pengujian sistem terhadap bacaan pengguna.

Tabel 2 Hasil Uji Respon Sistem

No	File Audio	Label Aktual	Prediksi	Confidence	Waktu Respon
1	samiiun1.wav	Benar	Benar	0.91	1.3 s
2	alasalah1.wav	Salah	Salah	0.87	1.2 s
3	alladi1.wav	Benar	Benar	0.93	1.4 s
4	alladi_salah.wav	Salah	Benar	0.78	1.5 s

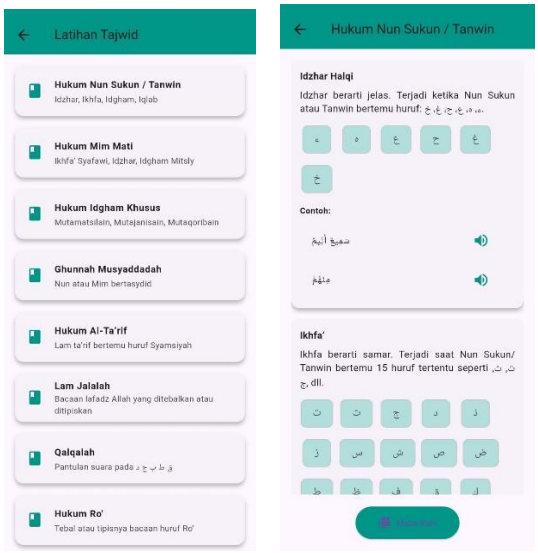
3.7. Implementasi Sistem

Pada tahapan ini, sistem dirancang melalui halaman awal. Kemudian diteruskan ke dalam halaman dashboard. Pengguna bisa memilih fitur yang disediakan, tampilan yang simple digunakan untuk mempermudah pengguna dalam menggunakan aplikasi, terdapat fitur waktu yang bisa menyesuaikan.

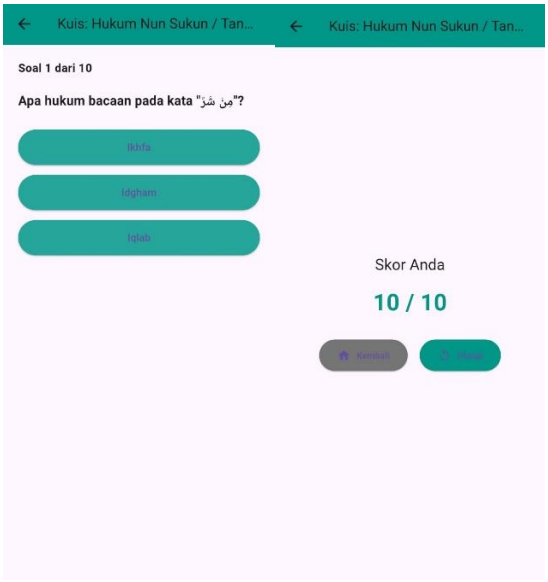


Gambar 4 Dashboard Aplikasi

Pada Halaman Materi Tajwid, terdapat beberapa list perkategori yang digunakan untuk mempermudah pengguna dalam mempelajari ilmu tajwid. Terdapat hukum bacaan dasar yang bisa dibaca dan dipelajari, dibuatkannya perpetak huruf dari setiap hukum bacaan yang berfungsi untuk mempermudah pengguna mengingat huruf hijaiyah disetiap hukum.

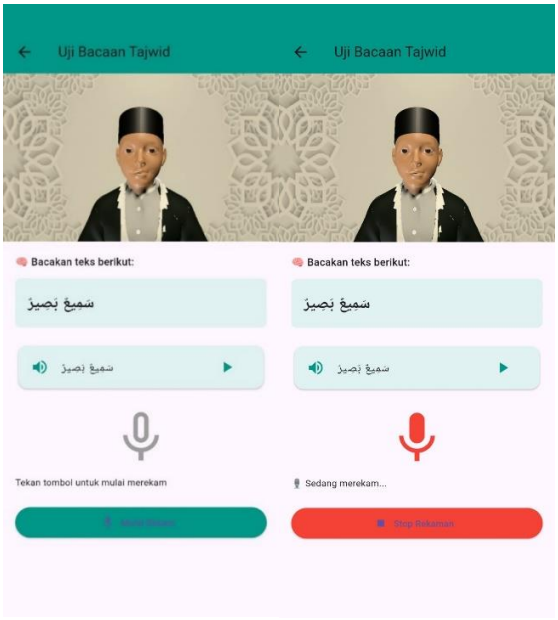


Gambar 5 Materi Tajwid



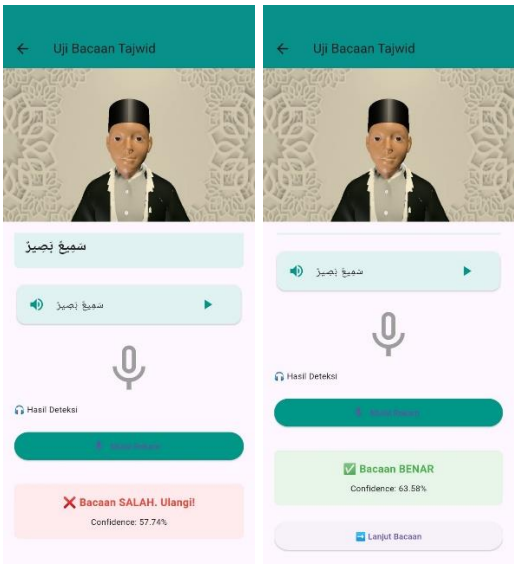
Gambar 6 Kuis

Dihalaman ini, pengguna dapat menjawab pertanyaan atau soal yang sudah disediakan, pengguna memilih jawaban dengan beberapa pilihan, dihalaman ini pengguna dapat mengetahui seberapa jauh pemahaman dengan nilai skor yang ditampilkan dalam aplikasi.



Gambar 7 Uji Bacaan

Pada fitur uji bacaan, pengguna dapat langsung membacakan bacaan sesuai teks, dan menekan kembali setelah selesai bacaan, dan dari itu sistem akan mengirim hasil bacaan berupa audio .wav ke server. Setelah sistem mendeteksi bacaan, sistem akan memberikan hasil berupa teks benar / salah dan feedback animasi berbicara beserta suara dari karakter secara realtime berdasarkan hasil uji bacaan.

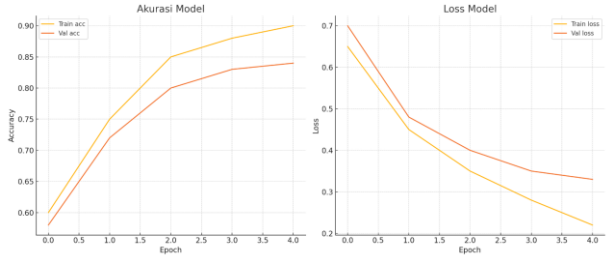


Gambar 8 Hasil Uji Jawaban

Hasil uji yang di tampilkan aplikasi terdapat 2 hasil dimana ketika hasilnya salah, pengguna harus mengulangi rekaman, dan ketika berhasil atau benar pengguna bisa melanjutkan bacaan.

3.8 Evaluasi Efektivitas Sistem

Evaluasi dilakukan menggunakan pendekatan black box testing dan Evaluasi Uji Akurasi model. Hasil blackbox menunjukkan bahwa seluruh fungsi sistem berjalan sesuai harapan, mulai dari fungsi fitur dan proses model sistem cnn-lstm.



Gambar 9 Evaluasi Tes Model

Dari uji training model terakhir mendapatkan hasil Akurasi Training Terakhir: 90.83% dan Akurasi Validasi Terakhir: 77.78%.

4. PENUTUP

4.1. Kesimpulan

Berdasarkan hasil implementasi dan pengujian yang telah dilakukan, dapat disimpulkan bahwa sistem aplikasi pembelajaran tajwid dan makharijul huruf berbasis speech Recognition telah berhasil dikembangkan dan diimplementasikan dengan baik. Model deteksi bacaan tajwid berbasis Speech Recognition yang dibangun menggunakan arsitektur CNN-LSTM mampu mengenali bacaan suara pengguna dan mengklasifikasikan secara baik antara bacaan benar dan salah. Hasil pengujian menunjukkan bahwa sistem bekerja secara stabil baik dalam aspek klasifikasi bacaan, pengolahan suara. Pengujian fungsional menggunakan metode Black Box Testing menunjukkan bahwa seluruh fitur berjalan sesuai dengan kebutuhan pengguna. Evaluasi menggunakan uji model memberikan hasil akurasi pelatihan sebesar 90,83% dan akurasi validasi sebesar 77,78% yang bisa dibilang cukup akurat.

4.2. Saran

Berdasarkan hasil penelitian, disarankan untuk mengembangkan sistem lenih lanjut dalam peningkatan kualitas dataset audio dengan variasi suara dan aksen pengguna agar model lebih general dan adaptif. Aplikasi ini disarankan untuk bisa dikembangkan kedalam aplikasi pada platform iOS dan Web untuk menjangkau lebih banyak pengguna. Kemudian menambahkannya fitur pelatihan mandiri atau latihan interaktif dengan peningkatan kesulitan secara bertahap agar sistem dapat memberikan kesan pembelajaran yang aplikatif dan edukatif.

5. DAFTAR PUSTAKA

[1] Ita Purnama Sari, I. ILMU TAJWID MELALUI METODE QIRO'ATI DALAM MEMBACA AL-QUR'AN (Doctoral dissertation, IAIN BENGKULU).

[2] Anggraini, N., Kurniawan, A., Wardhani, L. K., & Hakiem, N. (2018). Speech recognition application for the speech impaired using the android-based google cloud speech API. Telkomnika (Telecommunication Computing Electronics and Control), 16(6), 2733–2739. https://doi.org/10.12928/TELKOMNIKA.v16i6.9638

[3] Muhamad, S. O., Akter, F., Gali, M., Sains, F., & Teknologi, D. (n.d.). klasifikasi hukum tajwid mad dalam surah al-fatihah menggunakan algoritma long short-term memory dengan ekstraksi fitur mel-frequency cepstral coefficient program studi teknik informatika.

[4] Diarsyah, M. G., & Setiawan, D. (2024). IMPLEMENTASI CNN-LSTM UNTUK MUSIC CAPTIONING. In Media Informatika (Vol. 23, Issue 1).

[5] Arham, A. Z. (2018). Klasifikasi ulasan buku menggunakan algoritma convolutional neural network-long short term memory (Doctoral dissertation, Institut Teknologi Sepuluh Nopember).